Separating arguments from conclusions: The mistaken role of effect size in educational policy research

Adrian Simpson
School of Education, Durham University

Biography
Adrian Simpson is the Principal of Josephine Butler College and Professor in Mathematics Education at Durham University. His research has focused on the school-university transition in mathematics, proof, abstraction and assessment, though much of his more recent work has examined the way educational research impacts on policy and decision making.

Abstract
Effect size is the basis of much evidence-based education policy making. In particular, it is assumed to measure the educational effectiveness of interventions. Policy is being driven by the influential work of John Hattie, the Education Endowment Foundation and others, which is grounded in this assumption. This article demonstrates the assumption is false and notes that, when criticized, proponents either attempt to inoculate themselves by listing (without checking) assumptions or use the specious reasoning that, however flawed their argument, no-one has disproved their conclusions.

Key words: Effect size; evidence-based policy; statistical assumptions

> *"1031 is prime: 31 is prime and clearly adding 1000 to a prime number changes nothing," says Professor Corncrake. "Now, I have heard many people complain about my work and I am aware of the critiques. Indeed, I use some of them in my teaching. But as yet, no-one has convincingly argued that 1031 is not prime."*

Of course, Professor Corncrake is wrong – not in the fact of 1031 being prime, but in the argument that adding 1000 to a prime number results in another prime number. Professor Corncrake's defence is that the conclusion has not been proved wrong; but it does not seem to distinguish between concerns about that conclusion and concerns about the method. The conclusion happens to be correct, but the method is invalid (it works for vanishingly few primes). The study of number theory and its applications in, for example, keeping financial transactions secure, would be severely compromised if people relied on Corncrake's argument or accepted his defence.

This relatively straightforward example shows how important it is to distinguish between a conclusion and the warrant for drawing that conclusion. This paper explores how a lack of careful examination of the warrants for arguments is potentially misleading teachers and

other educational policy makers when they use the 'evidence based education' literature. It further examines the nature of the defences drawn by those who wish to promote an erroneous interpretation of the key measure used in this literature: effect size.

The Effect Size[1] in Education

The idea of using the effect size to compare educational interventions has a long history, but appears to have come to world-wide prominence through the highly influential book "Visible Learning" by John Hattie (Hattie, 2009). The work has led to his being called "possibly the world's most influential education academic" (Evans, 2012) and, according to Google Scholar (at the time of writing), this book has been cited nearly 10000 times.

Hattie (2009) draws what he calls 'bold speculations' by rank ordering dozens of forms of educational intervention - from 'feedback' towards the top and 'student control over learning' towards the bottom - arguing that the rank ordering is based on a "unidimensional continuum … that can allow various effects on achievement to be positioned as they relate to each other" (p.8). That 'unidimensional continuum' is based on the notion of effect size.

This is not the only influential work which encourages policy decisions to be grounded on effect size comparisons. Marzano (1998) undertook an earlier meta-meta-analysis using similar methods, and the notion of rank ordering interventions on the basis of effect size has gone on to be influential in many countries and in many subfields of education, resulting in a large number of publications and websites listing interventions in the order of their effect sizes as a proxy for their educational effectiveness. For example, in the UK, the Education Endowment Foundation (EEF) have developed a meta-meta-analysis toolkit, whose methods are based on effect size, as outlined in Higgins & Katsapataki (2016). This toolkit is even recommended as a source for policy making by the UK government (Department for Education, 2016), with two-thirds of school leaders saying they use it (EEF, 2016). Effect size ranking lists for subfields are also being produced, for example Schneider & Preckel (2017) give a rank ordering for higher education.

Eacott (2017) argues that Hattie's work has become the "dominant feature in contemporary educational leadership rhetoric in Australia" (p.414) and it can be argued that these similar approaches lie behind the key drivers of educational policy in other countries (Simpson, 2017). Thus far, the critiques of these approaches to combining studies and extracting policy proposals has focused on the difficulties with aggregating them in meta-analyses and meta-meta-analyses (Eysenck, 1984; Berk, 2007; Higgins and Simpson, 2011). This paper discusses the nature of what is being combined – the effect size – explores whether the assumptions made about what effect size represents hold and considers the nature of the defences used by proponents in the face of problems with their arguments and assumptions.

The basis of the argument

[1] Throughout this paper, 'effect size' will stand for 'standardised effect size' (as it does in the literature cited). While not exempt from all critique, raw effect size (where, for example, differences in scores are not scaled by a measure of variance) is not subject to some of the most serious problems noted here.

Each of the meta-meta-analytic rank orderings is based on comparing aggregated effect sizes. For the simplest of controlled trials, the effect size is defined as the difference between the mean score for the intervention group and the mean score for the control group, divided by some measure of the spread of those scores (such as the pooled standard deviation of the scores of the two groups). In meta-analyses, the effect sizes for a set of trials are averaged to purportedly give a more accurate estimate of the effect size of the intervention. In meta-meta-analyses, meta-analyses are still further combined under broader headings (such as 'homework' or 'teacher training') to provide an estimate of the effect size for classes of interventions. Finally, these broad areas are compared to provide the rank ordering, distinguishing "what are and what are not key influences on achievement" (Hattie, 2009, p15).

Fundamental to this work is the idea that effect size is a reasonable measure of the 'influence on achievement', that higher effect sizes are associated with larger influences and that interventions with similar educational influence have similar effect sizes.

Unfortunately, this is not the case.

Effect size is a measure of the trial as a whole; it is not a measure of the intervention. It tells us something about how clear the difference is between the intervention group and the control group on the outcome measure, not how educationally important or influential the intervention is. Simpson (2017) notes that effect size can be grossly affected by simple changes to any of the fundamental elements of a trial (the population from which the sample is selected, the activities undertaken by the control group, the chosen measure etc.). A simple thought experiment immediately shows that effect size cannot be taken as a valid measure of educational importance since it fails the basic validity requirement that the same intervention (with the same sample and control activity) should not result in very different effect sizes.

The Effect Size of a Thought Experiment

Imagine a sample of pupils drawn independently at random from some target population of, say, English speakers. That sample is randomly split into two groups. One group is told only one isolated fact - that the English translation of the Hungarian word 'oktatás' is 'education' - the other group gets no information at all. Both groups are then immediately given a test requiring them to translate ten Hungarian words into English.

The test contains the word 'oktatás' and nine other unfamiliar Hungarian words. Assuming that none of the pupils have any previous knowledge of Hungarian, the experimental group will likely score a mean of 1 out of 10 with very small standard deviation (almost everyone getting 'oktatás' and only 'oktatás' correct); the control group will likely score a mean of 0 out of 10 with a very small standard deviation (almost everyone getting no words correct). The effect size would be arbitrarily large: a difference of 1 divided by a standard deviation of near to zero (give or take the occasional lucky guess). In Hattie's (2009) terms, this would make the teaching of a single word of Hungarian the largest influence on achievement ever recorded. The EEF, who try to give meaning to the rather abstract notion of an effect size by scaling it to somehow represent 'months' progress', would have to declare this as the equivalent of many decades of additional progress in school (Higgins & Katsipataki, 2016).

However, if effect size is to work in the way required for producing ranked lists of effective educational interventions, it must have the property that very similar interventions have very similar effect sizes and, in particular, identical interventions (on identical samples and with other experimental design features fixed) have to have near identical effect sizes.

This is not the case: simple modifications to some aspects of the test in the thought experiment (without changing the intervention, sample or any other aspect of the trial) can drastically change the effect size:

a) If the test does not include the word 'oktatás', we would expect the effect size to be close to zero (everyone in both groups will likely score zero and any lucky guessing would be balanced in expectation between groups).
b) If the test contains 'oktatás' and one Hungarian word sufficiently similar to an English word that it might be guessed correctly by about half of the students (e.g. 'mobil'), the experimental group will still most likely score 1 mark more than the control (averaging 1.5 to the control group's 0.5) but the standard deviation is no longer effectively zero, it is around 0.25 (from the variation in guessing). So, the effect size will be around 4.
c) If the original set of test words were used in a multiple choice test with four options, then some answers will be entirely randomly guessed correctly (by both groups). The experimental group would be near perfect on 'oktatás', while the control group would be no better than chance on this. The control group would be expected to score around 2.5 marks, with a standard deviation which works out to around 1.4; the experimental group would get around 3.25 marks (one question almost certainly right and a quarter of the remaining 9 randomly correct), with a standard deviation of around 1.3 (there is slightly less variance as they are guessing on fewer questions). The effect size ends up being around 0.6.
d) If the multiple choice test envisaged in (c) had 20 questions (one being 'oktatás' and none of the others being familiar to anyone) instead of 10, the effect size ends up being around 0.4: the mean difference remains the same, but the standard deviation is multiplied by around 1.4. With 40 questions, the effect size is around 0.3, and so on.

In each of these examples the educational intervention (and control activity and population and sample size etc.) is exactly the same. In the thought experiment, the educational influence of the intervention has to be considered to be the same, as the educational outcome is the same - one group of people learn a single word of Hungarian more than the other group - yet the effect size varies from 0 to 0.4, to 0.6, to 4, to infinity.

Indeed, the choice of test might even lead to a negative effect size: imagine that the multiple choice test did not ask for the translation of 'oktatás' but 'okádás', giving the options 'pelican', 'teacher', 'vomit' and 'purple'. It is easy to imagine that the control group would distribute their answers evenly across the four options, but that the intervention group would disproportionately select the incorrect answer 'teacher', making the reasonable - if mistaken - assumption that words which seem similar have similar meanings, particularly in the context of the immediately preceding intervention.

This simple thought experiment shows that effect size cannot be a proxy for educational effectiveness or influence: exactly the same intervention, evaluated slightly differently, can result in effect sizes with widely different magnitudes and even potentially different signs. Moreover, altering the measure is not the only way to make very large changes in resulting effect size: in more complex interventions choosing a different control activity or reducing noise by selecting a more homogenous sample can make large differences even though the treatment received by the intervention group is identical.

One might question whether this thought experiment reflects real experience. On the one hand, the thought experiment is eminently implementable; on the other hand, there is ample evidence in the literature that the same intervention, on the same sample, with all other design features (except the test) held constant can lead to radically different effect sizes, showing that we cannot equate the effect size with the educational influence of the intervention.

The evaluation of a phonics intervention used the 'New Group Reading Test' (NGRT), a single word reading test (SWRT) and a Phonics Assessment Battery (PhAB) as outcome measures (Merrell and Kasim, 2015). For otherwise the same experiment (same intervention on the same sample, with the same control activities) they reported effect sizes 0.43, 0.38 and 0.23 respectively. While the evaluation team preselected one (NGRT) as their primary measure, a different evaluation of the same experiment could make a different choice and obtain a very different effect size.

The REACH targeted reading support programme evaluation (Sibieta, 2016) used the NGRT as the primary outcome and a reading comprehension score and a reading accuracy score as secondary outcomes. They found effect sizes of 0.33; -0.08 and 0.17 respectively. While the evaluators had pre-selected NGRT as the primary outcome, different evaluators looking at exactly the same intervention with other measures would have drawn quite different conclusions. In the 'months' progress' metric of the EEF we would be forced to conclude that the same intervention led to 4 months' progress, 2 months' progress and 1 month's regress.

The Nuffield Early Language Intervention evaluation (Sibieta, Kotecha and Skipp, 2016) involved a primary outcome (which was the composite of four measures) and a secondary outcome (the composite of three other measures). The effect sizes were 0.27 and 0.06 respectively. Again, the evaluators pre-selected a measure to count as 'the effect size', but different evaluators of the same intervention (in the same sample, with the same control activities etc.) could have chosen differently and drawn very different conclusions.

These real examples will, of course, be influenced by many other factors (not least measurement errors, differential attrition etc.) but alongside the thought experiment they show that the researchers' design decision to select (or develop) a test has a radical impact on effect size. Since otherwise identical experiments (the same intervention, sample, control group activity etc.) leads to radically different effect sizes depending of the researchers' selection of design features, it cannot be said to be a measure of the educational importance of an intervention. Larger effect sizes are not necessarily indicators

of increased educational influence: they may be due to test selections or other design features.

What does effect size actually measure?

Effect size is a measure of the *clarity* of the experimental difference between an intervention and a control group and it will be influenced by the design choices made by researchers. Indeed, as they seek increased statistical power (that is, their chance of finding a difference between groups, should it exist), any decision which does so (apart from increasing sample size or choice of statistical test) increases effect size. Indeed, for a fixed sample size, choice of statistical test and significance level, power and effect size are effectively the same thing: as in figure 1, they are simple continuous transformations of one another.

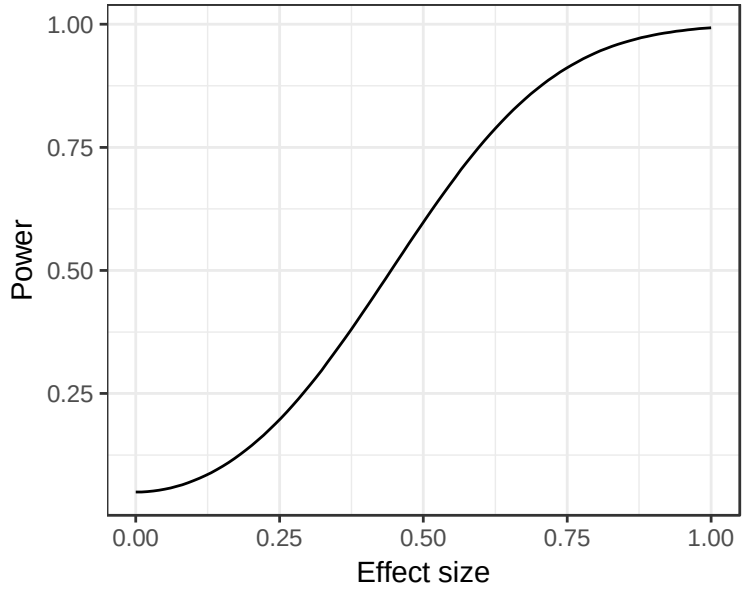

Figure 1. The relationship between power and effect size for a fixed sample size (in this case N=80), significance level (α=0.05) and statistical test (a two sample, two sided t-test).

As above, researchers can choose different tests (or even simply different lengths of a test). Or they can choose more or less restricted populations from which to draw their sample. They can have the control groups undertake different forms of alternative intervention. In each case, without changing the intervention (and, therefore, without altering its educational importance), they can vary the resulting effect size. Put another way – the enemy of clarity being *noise* – anything which acts to increase the noise of an experiment (multiple choice tests instead of open responses; heterogeneous samples; lower intervention dosage; increased interval between intervention and testing etc.) will decrease effect size.

Real educational environments are very noisy places (in many senses, but particularly in the statistical sense): the job of a researcher is often to amplify the signal so that it stands out from the noise. That is, they will make design decisions which increase effect size, within practical and ethical constraints.

Thus, even though the argument here shows that effect size should have no role in educational policy, it does have a technical role in research. It can tell future researchers using similar tests (and samples and control activities) something about the noisiness they might expect, allowing them to adjust sample size, reduce heterogeneity of the sample, select and modify measures, increase contrast with the control etc. as appropriate. The problem comes when, in mistakenly associating larger effect size with increased educational effectiveness, policy makers select between interventions. This may happen most often with meta-analysis and meta-meta-analysis[2], but happens also when comparing individual studies (e.g. Gorard, Siddiqui & See 2017).

With the REACH intervention, rather than concluding that the experiment was relatively clear when measured on the NGRT and rather unclear when measured on reading comprehension, the identification of effect size with educational influence would mean we would have to conclude the same intervention was both quite influential and rather non-influential. This makes little sense. Similarly, for the Nuffield Early language intervention: the mistaken identification of effect size with importance would lead to contradictory conclusions while the correct identification of effect size as experimental clarity would allow us to note that the difference between treatments on the sample was unclear when measured on the second test and rather clearer when measured on the first.

Whether the intervention is influential or not is a matter for professional judgement in context. It cannot be abandoned in favour of a numerical scale of experimental clarity, which is dependent on design features such as test selection, control activity and sample homogeneity.

However, even in the legitimate research use of effect size as a measure of experimental clarity, researchers' abilities to make those design choices can be more or less limited: depending on the context, researchers may struggle to remove noise from signal in their experiment. In simple, direct instruction contexts (such as the example in the thought experiment), researchers may have free rein in their choice of population, measure and control group activity. However, less direct influences may allow for less opportunity to increase experimental power by adjusting these factors. In behavioural intervention studies, it is likely to be unethical to use a 'no behavioural intervention' control and one is likely to use a standardised measure of achievement rather than a researcher designed one[3]. In the

[2] The argument from some meta-analysts is that all of the different design influences will 'wash out' when a large number of studies gets combined. But this relies on the obviously absurd assumption that studies are drawn independently and at random with the same distribution from a population of those design decisions: control activities, intervention-testing intervals, dosages etc. As if, for example, tests used in studies are selected at random from a population of possible tests so that the number of open answer vs multiple choice tests 'balances out' in some way. See Berk (2007, 2011).

[3] Cheung & Slavin (2016) noted that, across their sample of studies, researcher designed measures resulted in effect sizes on average twice the size of those from studies with independently designed measures. The reason for this follows directly from the thought experiment here: researchers can (and do) reduce noise by designing tests which target only the impact of the intervention; standardised tests will not do so. Though, of course,

earlier examples of real projects which give different effect sizes, the restriction on the choice of tests is imposed by the funding body.

The appearance of direct instruction interventions like feedback and meta-cognition at the top of meta-meta-analytic tables should be taken only as evidence that researchers have found it easier to conduct less noisy experiments in these areas compared to behaviour interventions or summer schools. It is a mistake to argue that it means feedback or meta-cognition are more influential educational interventions.

A Fundamental Category Error

Mistaking effect size for a measure of educational importance or influence is, then, a category error. But a category error can be more or less serious. A personal shopper who selects trousers for their customers on the basis of their height rather than their inside leg measurement is making a category error, but there is a close correlation between the two and one might expect the trousers to fit on a few occasions. Making the selection based on the length of the customer's surname is a larger category error: there is no obvious correlation between the two and the fit would not be expected to be better than random (even though the process is systematic).

There is little obvious reason to believe that educational influence is even closely correlated to study clarity (or lack of noise). If effect size was somehow a measure of educational influence, it would be expected to be independent of researchers' design choices (which it is clearly not); as a measure of experimental clarity, we would expect it to be dependent on design (in the ways shown above). In the thought experiment, the educational importance of teaching a single word of Hungarian would appear very low, but with such a direct form of instruction, it is easy for researchers to select a measure with as high a signal-to-noise ratio as they like.

Indeed, 'educational influence' is surely a context bound concept: teaching a single word of Hungarian would otherwise seem trivial, but knowing that word as you search among a Bucharest university campus for the building holding the education conference might make it appear more important. Or, learning this one word in the context of starting to learn a new language might make learning it mildly educationally important while learning it as a one-off in the context of a psychology memory study might make it educationally trivial.

Equally, in another context, an educationally influential effect might (in the presence of much noise) result in a very small effect size. A programme which very effectively teaches people lifesaving CPR (cardiopulmonary resuscitation) might result in a smaller effect size if tested on a widely heterogenous sample (with people across a wide range of pre-existing knowledge of CPR); if measured using a standardised first aid test (rather than a CPR specific test); if compared to a somewhat less effective teaching method (rather than no CPR teaching whatsoever). Teaching people CPR – in the right context – may be educationally very important (literally lifesaving) and this intervention could be a very effective way of delivering it, no matter that the study results in a small effect size. If we read effect size as a

---

researchers can (and do) still *select* standardised tests to reduce noise and amplify the signal.

measure of practical significance or educational influence, rather than a measure of experimental clarity, we might dismiss an approach which could be very important indeed.

Some might think that this criticism of the identification of effect size with educational influence is merely technical or only applies to meta-analysis (or meta-meta-analysis) and technical fixes will make the problems disappear. While combining this flawed identification with the conceptual flaws in meta-analysis (Berk, 2007) magnifies the problem, the criticism is fundamental. If exactly the same intervention on the same sample, compared to the same control activity leads to radically different numerical outcomes, those numbers cannot be measures of educational effectiveness.

It should be noted that if – and only if – the same intervention is undertaken in multiple studies with the same sample, compared to the same control activity, measured on the same outcome (and using the same study design and statistical analysis), then one can combine the effect sizes from those studies. Moreover, if different interventions are evaluated on the same sample, against the same control activity, with the same outcome measure, design and statistical analysis, then relative effect size may tell us which has been more effective. Of course, in such cases, one might just as well work with raw effect size (Baguley, 2009). In reality, direct replications and comparisons of this sort do not happen and are not the basis of 'evidence based policy'.

Defending against the Criticisms

The failure of effect size to act as a proxy for educational effectiveness should lead us to question previous policy decisions based on comparing them or combining them. Yet responses to demonstrations which refute the logic of these arguments or which show that meta-analyses and meta-meta-analyses make assumptions which are obviously false (Berk, 2007; Bergeron, 2017; Simpson, 2017) take predictable defensive forms. Freedman (2009) listed forms of defence offered by those whose statistical arguments have been exposed as flawed:

> The assumptions are reasonable.
> The assumptions don't matter.
> The assumptions are conservative.
> You can't prove the assumptions are wrong. …
> We're only doing what everybody else does.
> Now we use more sophisticated techniques.
> If we don't do it, someone else will.
> What would you do?
> The decision-maker has to be better off with us than without us. …
> Where's the harm? (p 212)

Many of these arguments are clearly visible in the defences offered by some of the key players in promoting effect size for setting policy. For example, the first three forms of defence are implicit when authors simply list the assumptions, without checking whether they hold. For example, Schneider and Preckel (2017) note, amongst other assumptions which need to hold for an educational meta-meta-analysis rank ordering to be meaningful: the need for treatment intensity in different studies to be comparable; the need for

different meta-analyses to use similar inclusion/exclusion criteria; the need to avoid meta-analyses which focus on teaching methods which might be disproportionately suited to particular content. Similarly, Higgins & Katsipataki (2016) list concerns about publication bias, failure to account for nesting or clustering of schools and comparability of implementation.

Berk (2007) analysed this defence strategy: the response to "the mismatch between a meta-analysis model and anything real …[is that] … the requisite assumptions are listed, but not defended. A list of the assumptions by itself apparently inoculates the meta-analysis against modelling errors" (p.264). Despite listing a few underlying assumptions, Schneider & Preckel (2017) and Higgins & Katsipataki (2016) do not check whether they hold in their own data, and *prima facie* these and many other assumptions necessary for valid arguments are not met. In particular, as shown here, simple measure validity (whether the measure corresponds to its proposed use) is violated: effect size is not a valid measure of educational influence.

Another of the defence strategies identified by Freedman (2009) is to claim that the results should stand because the critics have not proved the assumptions wrong. Just as Professor Corncrake argues that no-one has shown that 1031 is not a prime number, many who continue to use effect size to stand for educational effectiveness seem unable to separate a flawed argument from the possibility that some of the conclusions may be spuriously correct (but may equally be wildly incorrect). For example, despite the criticisms of his warrant, Hattie (2017) argues that his headline conclusions have not yet been falsified:

> "there have been critics of some of the data, re-interpretations from some of the meta-analyses, but so far no critique (that I know about) about the 'bold speculations' about seeing learning through the eyes of the students, enabling students to become their own teachers, the value of success criteria, the premise of 'Know thy impact' in its many forms, the role of relationships and trust to allow for error and misconceptions to come to the fore, about the power of giving and receiving feedback, the focus on learning strategies in the context of the subject domain, making schools inviting places for students to come and learn, and focusing on the learning lives of students." [p.428]

We do not yet know whether feedback is generally a better intervention than homework or whether homework is better than teacher training – arguably, these are not even well formed educational questions – so we are not able to falsify (nor confirm) Hattie's "bold speculations", but we can note that the invalid form of argument that he uses (along with the EEF; Marzano, 1998; Schneider & Preckel, 2017 and others) does not allow their conclusions to be drawn. It certainly means we should not be basing policy on them.

Freedman's (2009) final form of defence is "where's the harm?": Forms of intervention are promoted as more effective when, in fact, the evidence merely indicates they are areas in which it may be easier to conduct clearer studies. This misidentification is leading policy, driving the use of scarce resources and causing major changes in teaching methods.

There's the harm.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British Journal of Psychology*, *100*(3), 603-617.

Bergeron, P. (2017). How to engage in pseudoscience with real data: a criticism of John Hattie's argument in Visible Learning from the perspective of a statistician, *McGill Journal of Education*, 52(1), 237-246.

Berk, R. (2007). Statistical inference and meta-analysis. *Journal of Experimental Criminology*, *3*(3), 247-270.

Berk, R. (2011). Evidence-based versus junk-based evaluation research: Some lessons from 35 years of the evaluation review. *Evaluation Review*, *35*(3), 191-203.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283-292.

Department for Education. 2016. *Educational Excellence Everywhere*. London: HMSO.

Eacott, S. (2017). School leadership and the cult of the guru: the neo-Taylorism of Hattie. *School Leadership & Management*, 1-14.

EEF (2016). *The EEF at Five*, London: Education Endowment Foundation

Evans, D. (2012) He's not the messiah, *Times Educational Supplement*, 14th September, 2012.

Eysenck, H.J. (1984). "Meta-analysis: an abuse of research integration" *Journal of Special Education* 18, 41-59.

Freedman, D. (2009) *Statistical Models: Theory and Practice*, Cambridge: Cambridge University Press.

Gorard, S., Siddiqui, N., & See, B. H. (2017). What works and what fails? Evidence from seven popular literacy 'catch-up' schemes for the transition to secondary school in England. *Research Papers in Education*, *32*(5), 626-648.

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Abingdon: Routledge.

Hattie, J. (2017). Educators are not uncritical believers of a cult figure. *School Leadership & Management*, *37*(4), 427-430.

Higgins, S., & Katsipataki, M. (2016). Communicating comparative findings from meta-analysis in educational research: some examples and suggestions. *International Journal of Research & Method in Education*, *39*(3), 237-254.

Higgins, S., & Simpson, A. (2011). Review of Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement, *British Journal of Educational Studies*, 59 (2), 197–201

Marzano, R. J. (1998). *A Theory-Based Meta-Analysis of Research on Instruction*. Aurora, Colarado: Mid-continent Regional Educational Laboratory.

Merrell, C. & Kasim, A. (2015) *Butterfly Phonics: Evaluation report and executive summary*. London: Education Endowment Foundation.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological bulletin*, *143*(6), 565.

Sibieta, L. (2016). *REACH: Evaluation report and executive summary.* London: Education Endowment Foundation.

Sibieta, L., Kotecha, M. and Skipp, A. (2016) *Nuffield Early Language Intervention: Evaluation report and executive summary.* London: Education Endowment Foundation

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*(4), 450-466.